Face modeling (part 2)

Jun-Yan Zhu

16-726 Learning-based Image Synthesis, Spring 2022

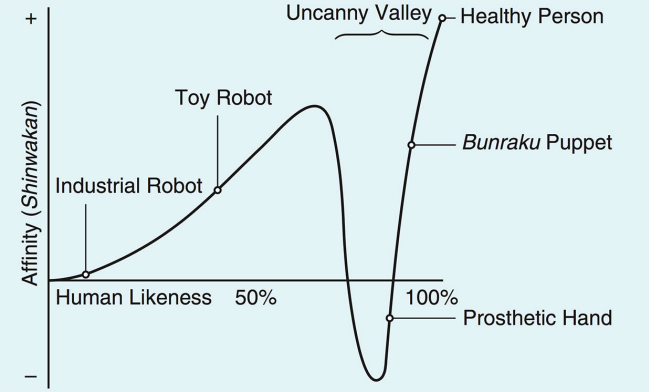
Slides adopted from Alexei A. Efros , Hyeongwoo Kim et al.,

Why Human Faces?

- Face is an important subject.
  - We are humans.
  - Many commercial applications.
- Lots of useful tools
  - 3D data: geometry-based synthesis.
  - 2D/3D Computer vision works for faces.

Is Face Modeling Easy/Hard?

- Face modeling is easy?
  - Plenty of aligned 3D face data.
  - 2D and 3D computer vision methods.
- Face modeling is hard?
  - Uncanny valley: Human eyes are extremely sensitive to any imperfections on faces.

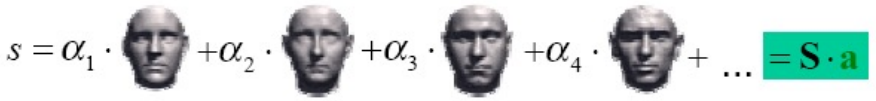


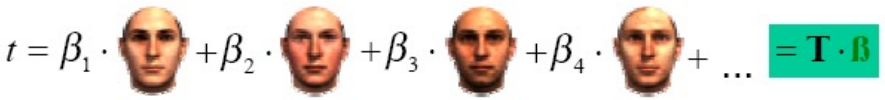
Masahiro Mori

The Morphable face model (3D + PCA)

Again, assuming that we have ***m*** such vector pairs in full correspondence, we can form new shapes $\mathbf{S}_{model}$ and new appearances $\mathbf{T}_{model}$ as:

$$\mathbf{S}_{model} = \sum_{i=1}^{m} a_i \mathbf{S}_i \qquad \mathbf{T}_{model} = \sum_{i=1}^{m} b_i \mathbf{T}_i$$





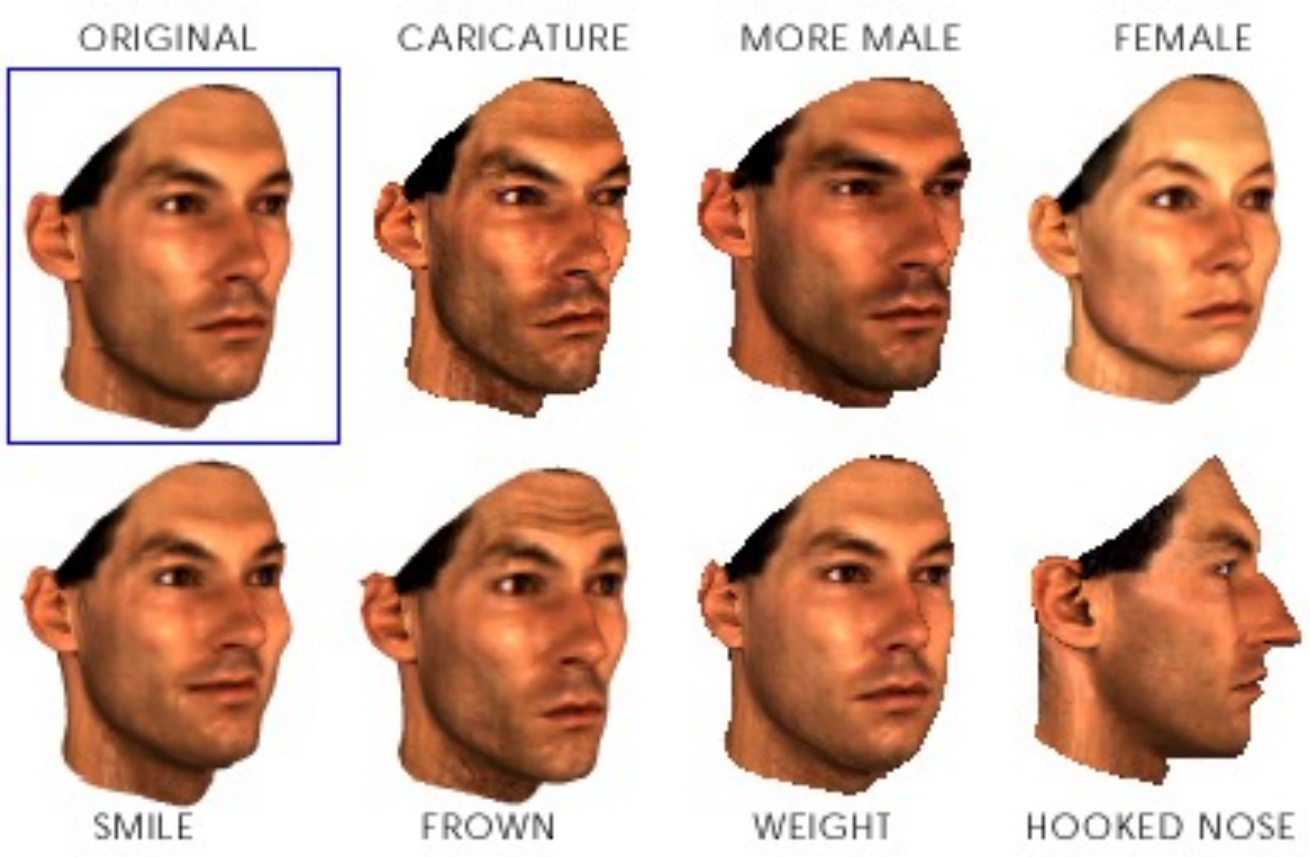
If number of basis faces ***m*** is large enough to span the face subspace then:

Any new face can be represented as a pair of vectors

$(\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ and $(\beta_1, \beta_2, \ldots, \beta_m)^T$!

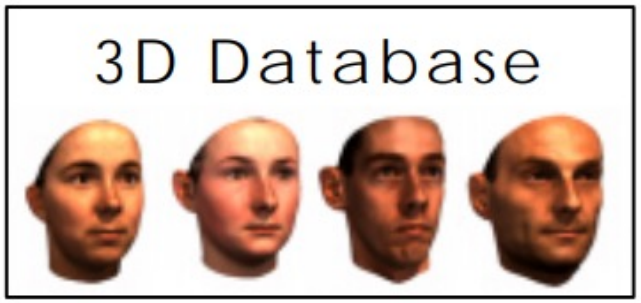
Blanz and Vetter, SIGGRAPH 1999

Using 3D Geometry: Blinz & Vetter, 1999



Blanz and Vetter, SIGGRAPH 1999

Using 3D Geometry: Blinz & Vetter, 1999

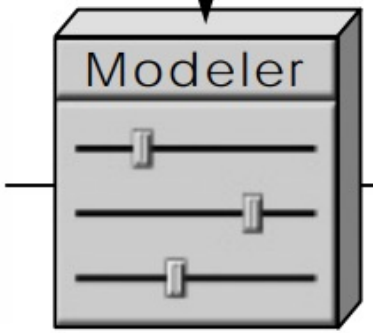


Morphable Face Model

$$\mathbf{T}_{model} = \sum_{i=1}^{m} b_i \mathbf{T}_i$$

2D Input

Face Analyzer



3D Output

$$E_I = \sum_{x,y} \| \mathbf{I}_{input}(x,y) - \mathbf{I}_{model}(x,y) \|^2 .$$

Input image

Phong illumination model

$s = \alpha_1 \cdot$  $+\alpha_2 \cdot$  $+\alpha_3 \cdot$  $+\alpha_4 \cdot$  $+ \ldots = \mathbf{S} \cdot \mathbf{a}$

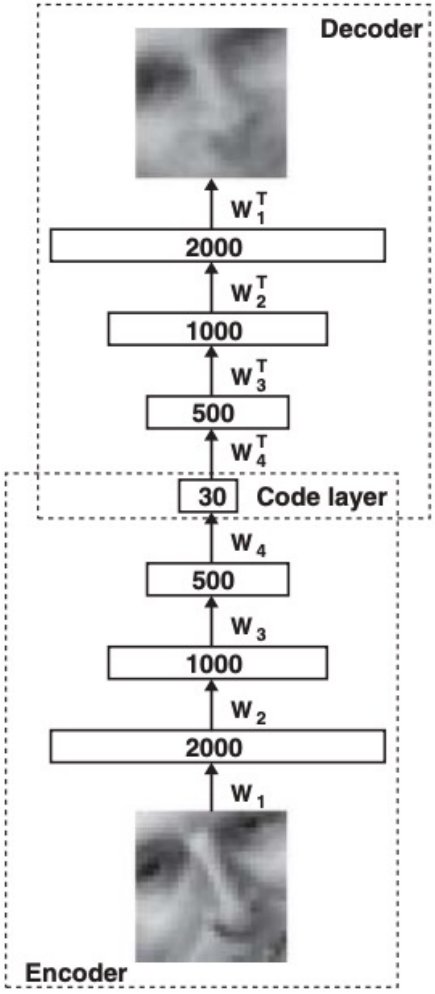
$t = \beta_1 \cdot$  $+\beta_2 \cdot$  $+\beta_3 \cdot$  $+\beta_4 \cdot$  $+ \ldots = \mathbf{T} \cdot \boldsymbol{\beta}$

Blanz and Vetter, SIGGRAPH 1999

How to Improve the results?

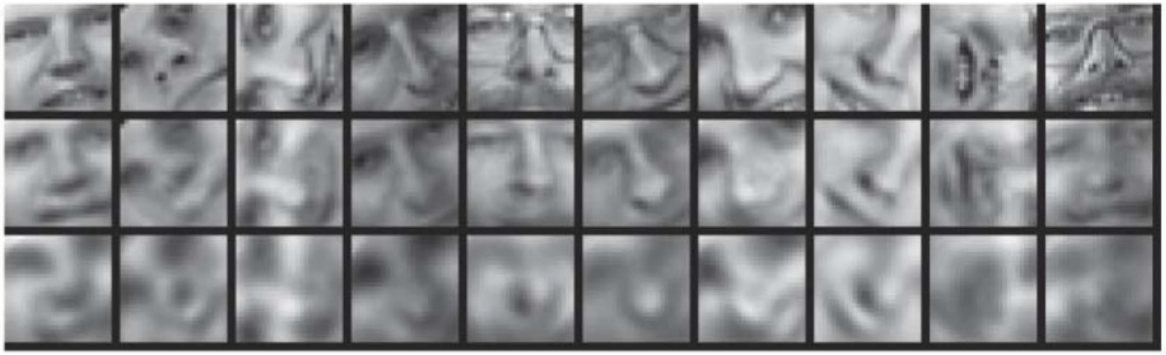
- Using Deep Learning?
- But how?
- Deep learning vision methods:
  - 2D/3D landmark detection
  - 3D pose estimation
  - Face shape reconstruction
- Deep learning graphics models
  - generative models
  - 3D-aware generative models

Autoencoder vs. PCA



Training objective: E encoder, G decoder/generator

$$\arg\min_{E,G} \mathbb{E}_x ||G(E(x)) - x||_2$$



Top: Input. Middle: Autoencoder. Bottom: PCA

GE Hinton, RR Salakhutdinov. Science 2006

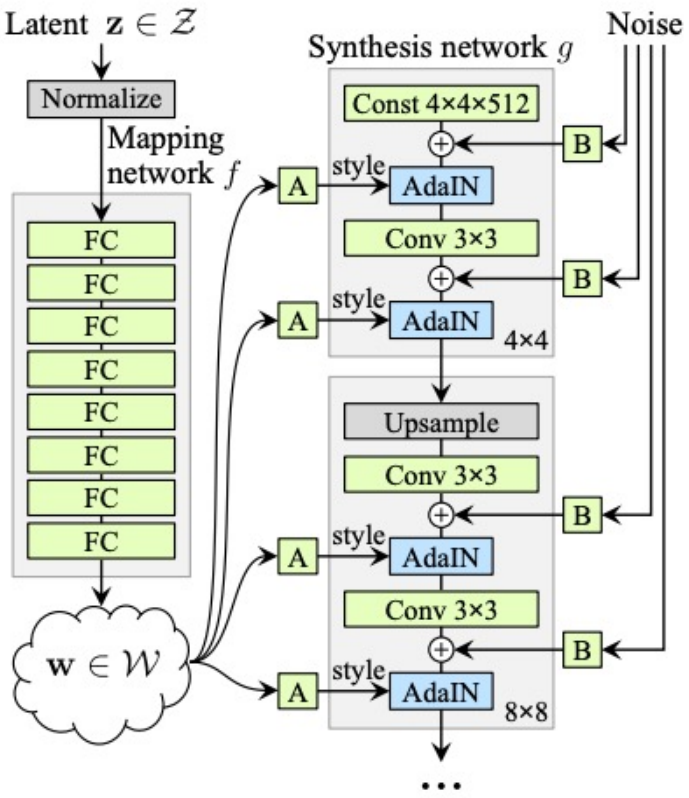
Deep learning method

PCA→ Generative Model

StyleGAN Face Results

StyleGAN [Karras et al., 2019]

Face Editing with GANs Projection



Optimizing the latent code

$$z^* = \arg\min_z \mathcal{L}(G(z), x)$$

Optimizing the style code

$$w^* = \arg\min_w \mathcal{L}(g(w), x)$$

Optimizing the extended style code

$$w_+^* = \arg\min_{w+} \mathcal{L}(g(w_+), x)$$

Image2StyleGAN [Abdal et al., 2019], StyleGAN2 [Karras et al., 2019]

11

Face Editing = latent space editing

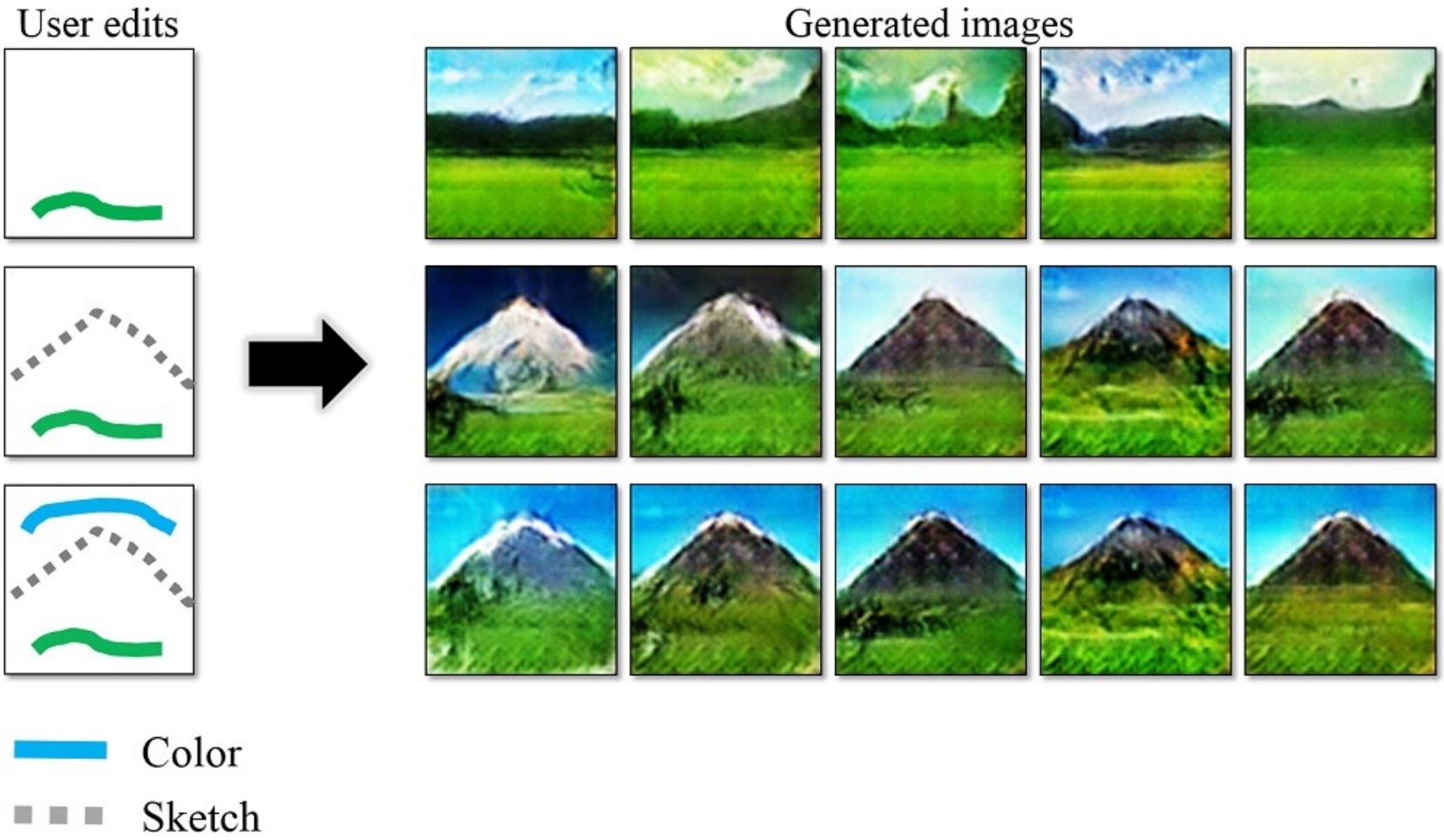
Interpolation between two faces in the w+ space

Image2StyleGAN [Abdal et al., 2019], StyleGAN2 [Karras et al., 2019]

12

Face Editing = latent space editing

Image Editing with GANs Projection



iGAN [Zhu et al., 2016]

Face Editing with GANs Projection

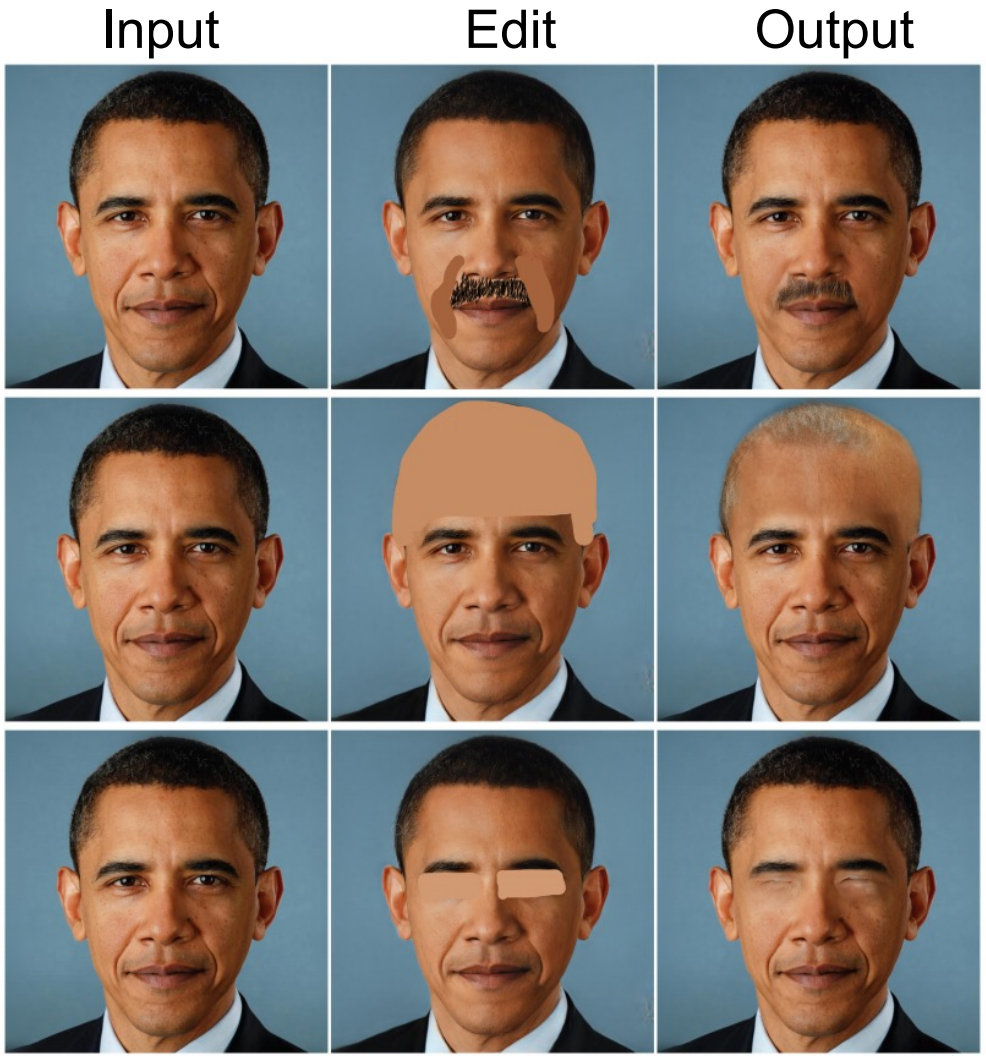


Image2StyleGAN++ [Abdal et al., 2020]

Face Editing with GANs Projection

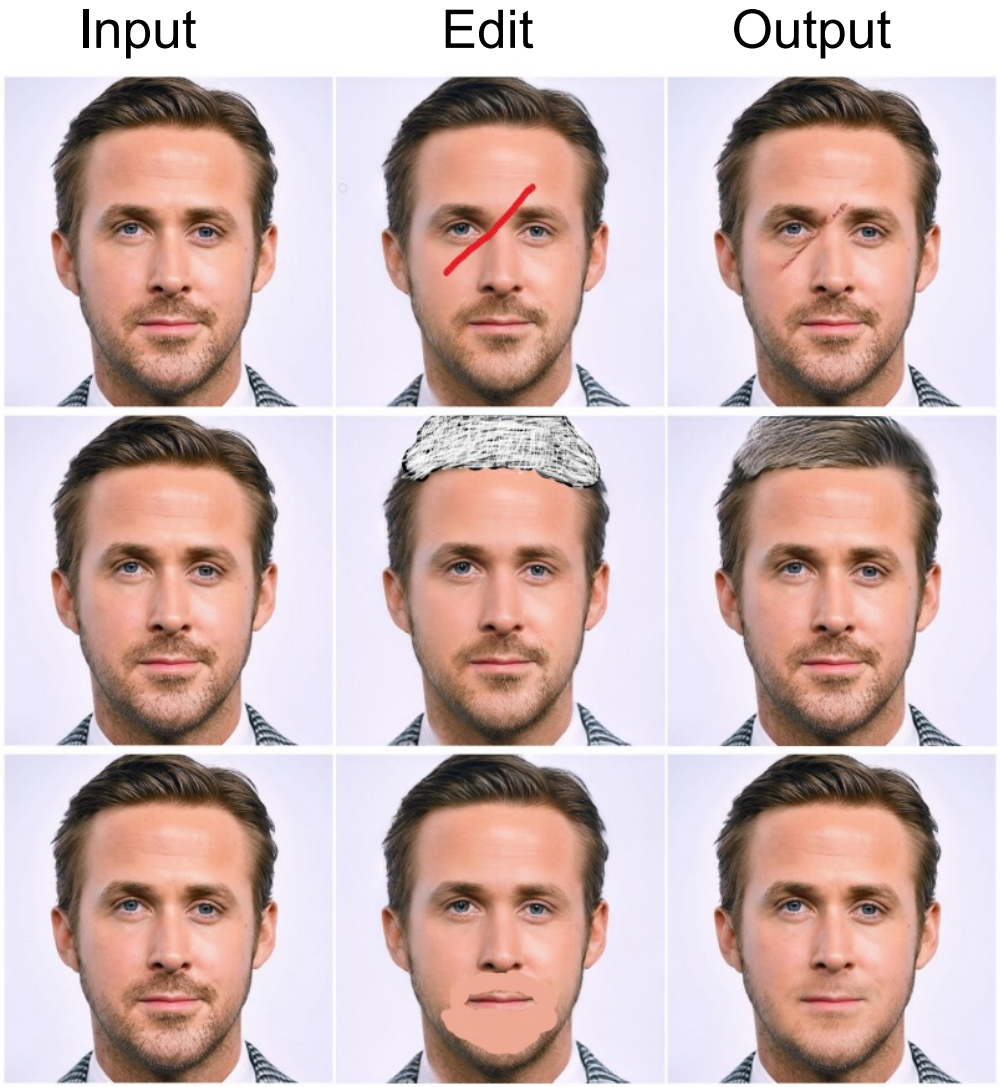
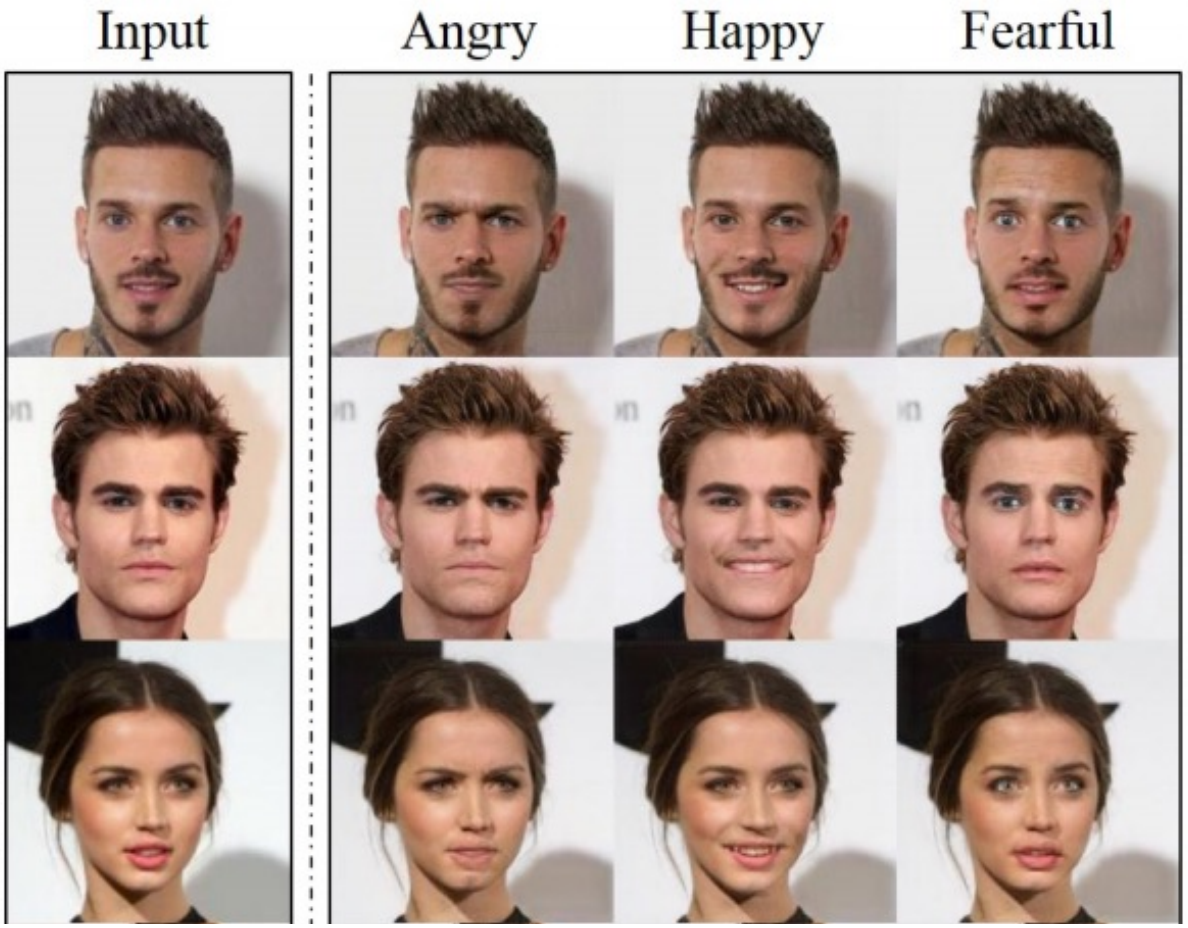


Image2StyleGAN++ [Abdal et al., 2020]

Deep learning method

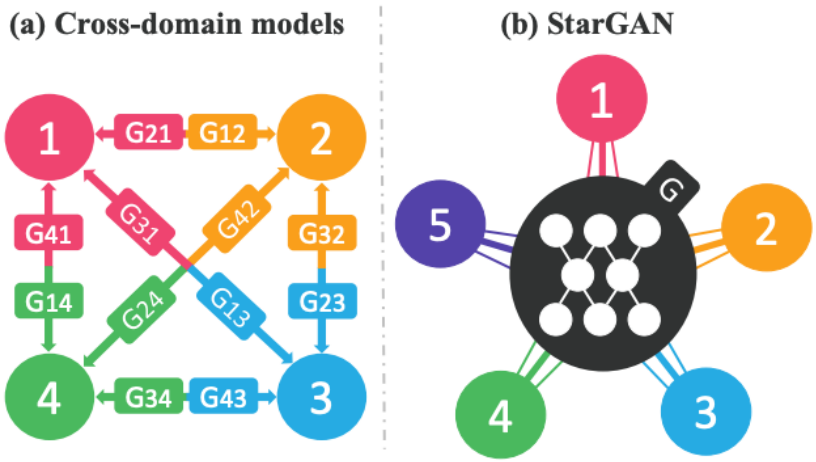
Image-to-Image Translation

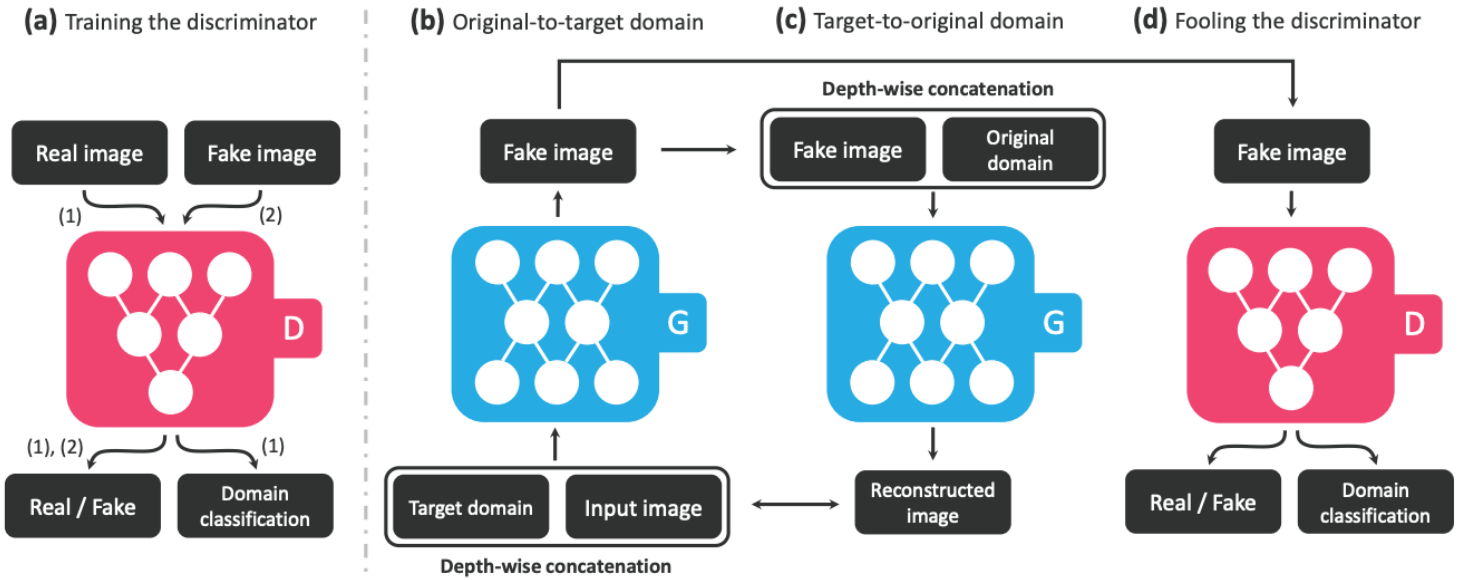
Face Translation with StarGAN



StarGAN [Choi et al., 2018]

Face Translation with StarGAN





StarGAN [Choi et al., 2018]

Face Translation with StarGAN v2

Multi-modal synthesis; supports a reference image

StarGAN v2 [Choi et al., 2020]

3D + Deep Learning

3D representation+ image-to-image

CGI Face Editing

Professional video

Video: © The Curious Case of Benjamin Button

CGI Face Editing

Personal video

Video: © https://www.youtube.com/watch?v=7Flvkn2quLY

Applications

Original video Pose editing Expression editing

- Editing of head pose, rotation, face expression and eye gaze
- Combination of model-based face capture and CNN

Video: courtesy of UK government
(Open Government Licence)

3D + CNN

Model-based face capture and reenactment

Garrido et al., ToG 2016

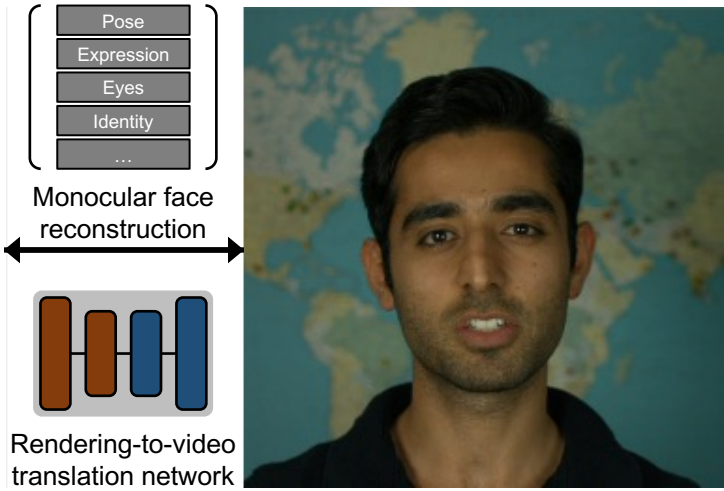
Kemelmacher-Shlizerman et al., ECCV 2010
Shi et al., ToG 2014
Suwajanakorn et al., ICCV 2015
Thies et al., CVPR 2016
Averbuch-Elor et al., ToG 2017
Thies et al., SIGGRAPH 2018

CNN-based methods

Karras et al., ICLR 2018

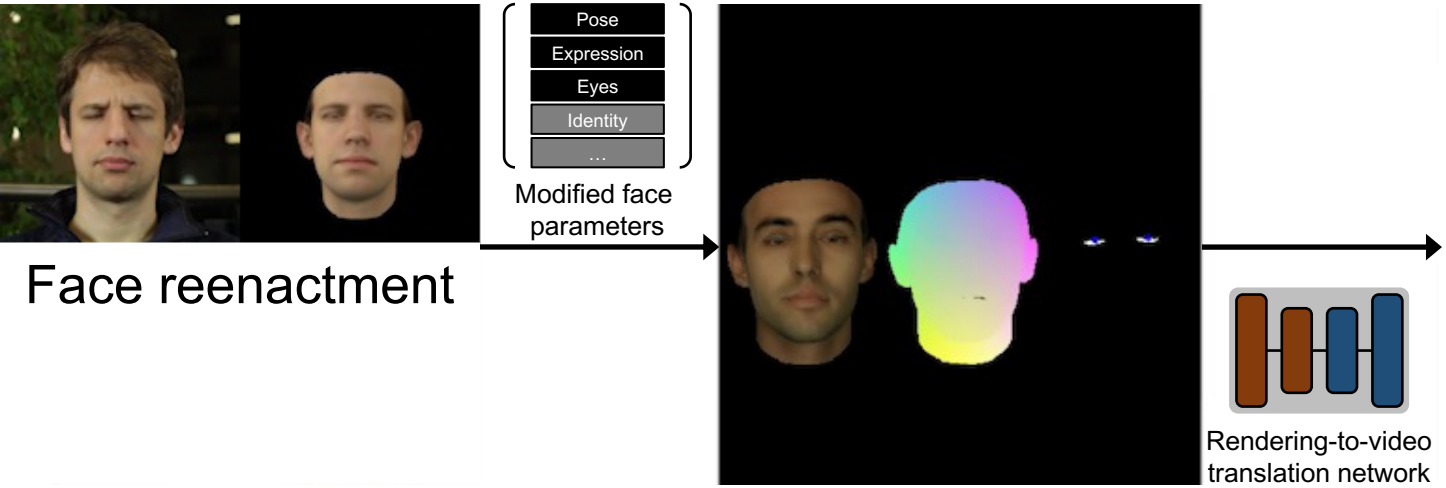
Goodfellow et al., NIPS 2014
Isola et al., CVPR 2017
Chen and Koltun, ICCV 2017
Tewari et al., ICCV 2017
Olszewski et al., ICCV 2018
Wang et al., CVPR 2018

Overview



Deep video Portrait [Hyeongwoo et al., SIGGRAPH 2018]

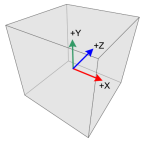
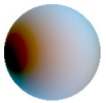
Overview



Deep video Portrait [Hyeongwoo et al., SIGGRAPH 2018]

Monocular 3D Face Reconstruction

- Parametric 3D face model

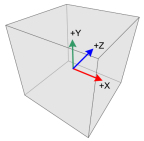
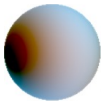
$p = ($  ,  ,  ,  $) \in \mathbb{R}^{257}$

Pose Expression Identity Lighting

$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

Monocular 3D Face Reconstruction

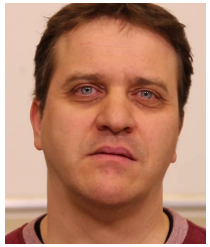
- Parametric 3D face model

$p = ($  ,  ,  ,  $) \in \mathbb{R}^{257}$

Pose Expression Identity Lighting

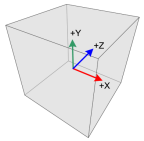
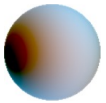
$$\min_{p} E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

$\| \quad - \quad \|_2^2$



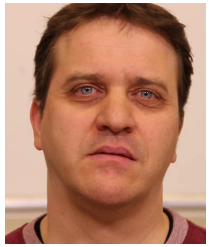
Monocular 3D Face Reconstruction

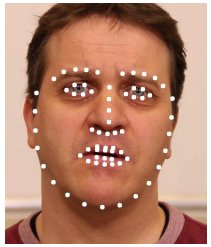
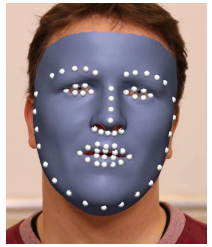
- Parametric 3D face model

$p = ($  ,  ,  ,  $) \in \mathbb{R}^{257}$

Pose Expression Identity Lighting

$$\min_{p} E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

$\| \quad - \quad \|_2^2$  

$\| \quad - \quad \|_2^2$  

Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\quad , \quad , \quad , \quad) \in \mathbb{R}^{257}$$

Pose Expression Identity Lighting

$$\min_{p} E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

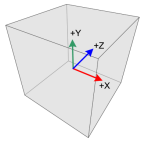
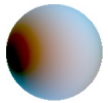
$$\| \quad - \quad \|_2^2 \qquad \| \quad - \quad \|_2^2$$

Statistical and temporal regularization

Garrido et al., ToG 2016

Monocular 3D Face Reconstruction

- Parametric 3D face model

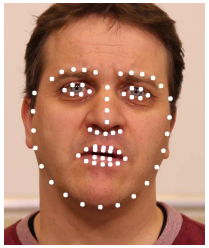
$p = ($  ,  ,  ,  $) \in \mathbb{R}^{257}$

Pose Expression Identity Lighting

$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

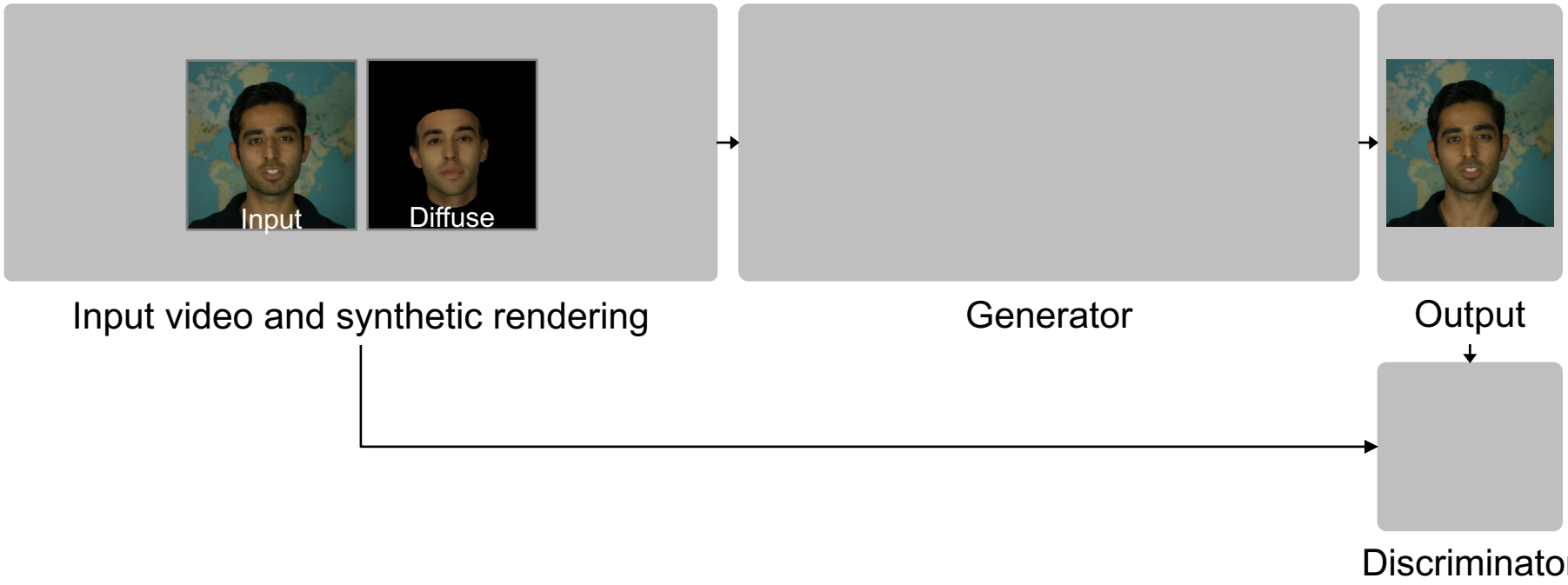
- Eye model

$e = ($  $) \in \mathbb{R}^{4}$



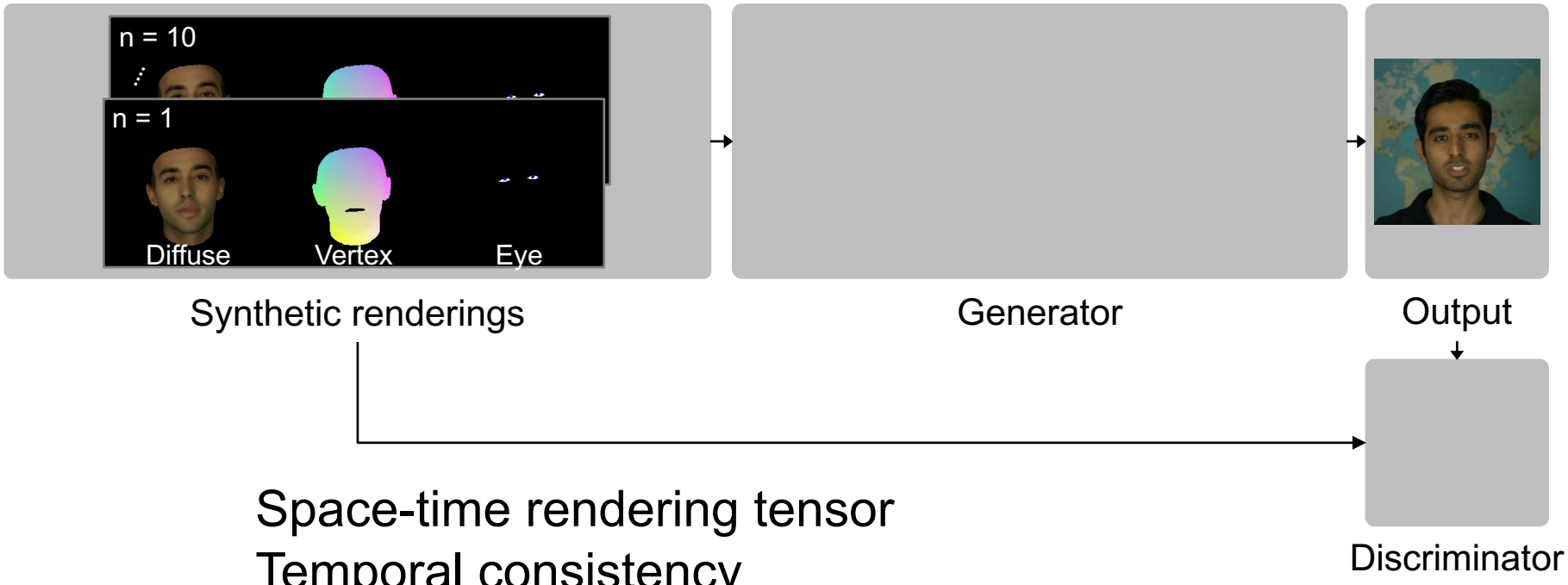
Saragih et al., FG 2011

Rendering-to-Video Translation Network

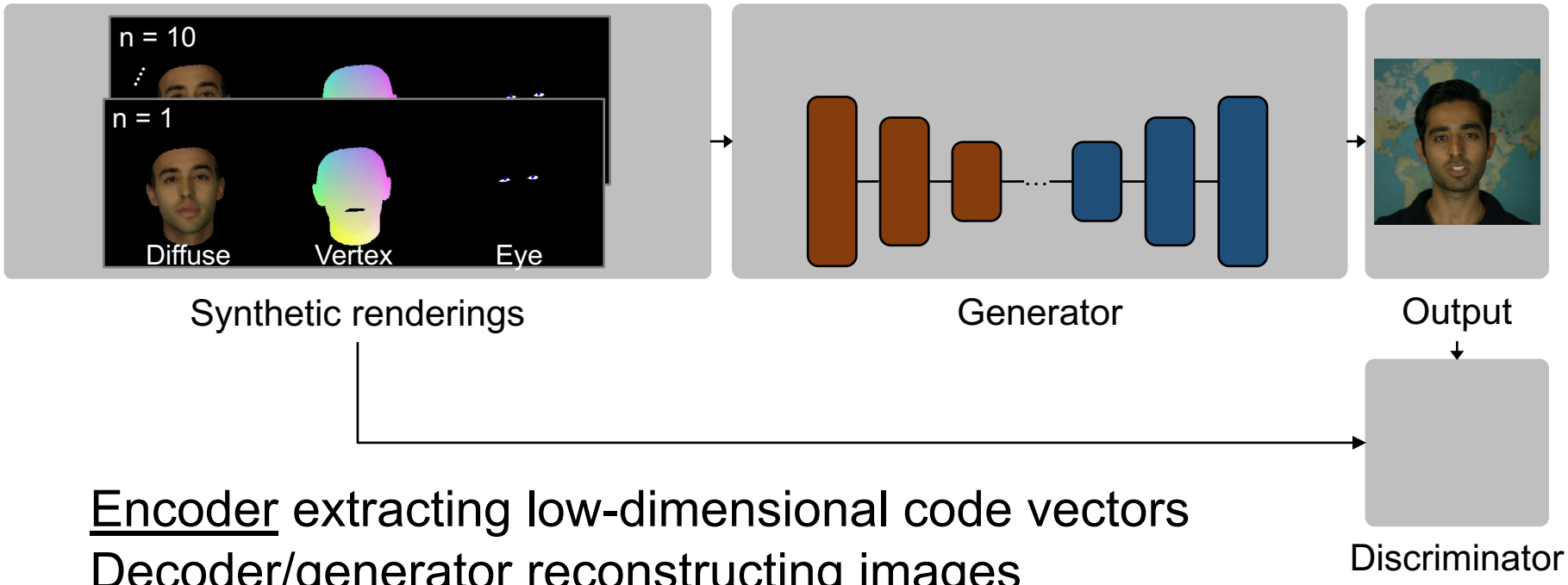


Pix2pix [Isola et al., ICCV 2017]

Rendering-to-Video Translation Network

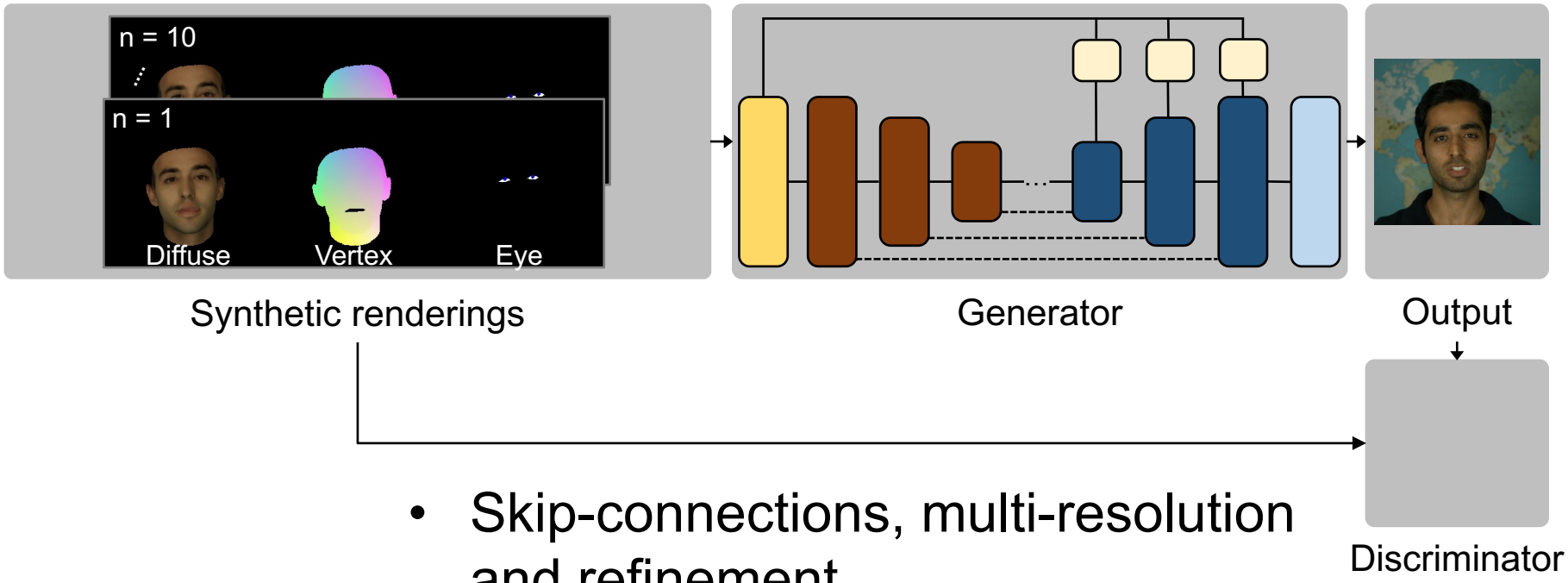


Rendering-to-Video Translation Network



Encoder extracting low-dimensional code vectors
Decoder/generator reconstructing images

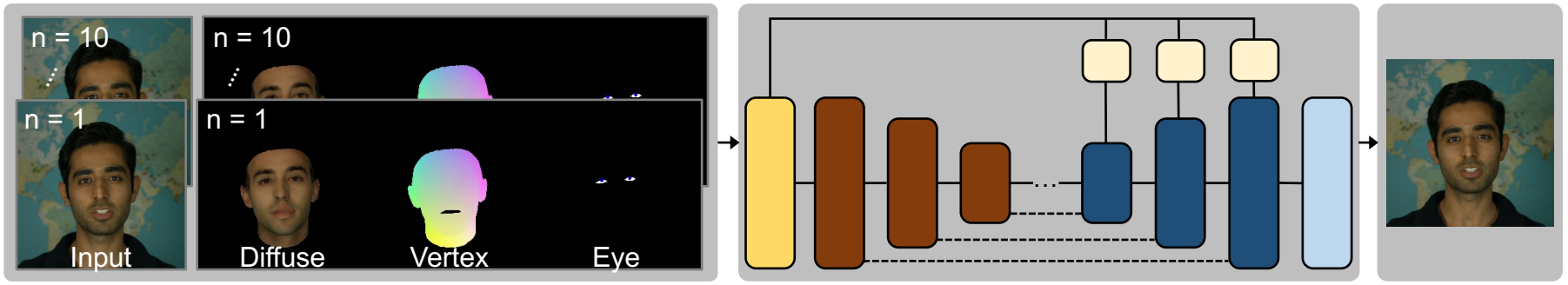
Rendering-to-Video Translation Network



- Skip-connections, multi-resolution and refinement
- Fine-scale details

U-Net [Ronneberger et al., MICCAI 2015]
CRN [Chen and Koltun, ICCV 2017]

Rendering-to-Video Translation Network



Training with adversarial and L_1 losses

GANs [Goodfellow et al. NIPS 2014]
Pix2pix [Isola et al. ICCV 2017]

Result: Facial Reenactment

Retargeting portraits videos from source to target

Source Result

Result: Facial Reenactment

Full reenactment of head pose, head rotation, face expression and eye gaze

Source | Result | Face2Face (Thies et al., 2016)

Result: Facial Reenactment

Source Target Result

Video: courtesy of the White House
(public domain)

Result: Visual Dubbing

Visual discomfort due to the discrepancy between video and audio tracks

Dubbing actor video Original video

Result: Visual Dubbing

Modification of mouth motion to match audio tracks

Dubbing actor video Dubbed video Garrido et al., 2015

Result: Interactive Editing

Pose Expression Shape

Approximately 9 fps

Result: Interactive Editing

YouTube videos 2× speed

Approximately 9 fps

Reagan video courtesy of NARA (public domain)
Obama video courtesy of the White House (public domain)
Wolf video courtesy of Tom Wolf (CC BY)

Result: Post-Production

Face reshaping Subtle expression editing

The Curious Case of Benjamin Button
video courtesy of Lola Visual Effects

Result: Pose Correction in Teleconferencing

Modification of head pose to match camera views

Setup | Camera view | Rotating up

Result: Multi-View Teleconferencing

Rotating up + side to side

Model-based video coding: 31 KB/s
h.264 (e.g., Skype): 192 KB/s

3D + Deep Learning

3D representation+ StyleGAN

StyleRig

Face reconstruction network as encoder + StyleGAN as generator

StyleRig: Rigging StyleGAN for 3D Control over Portrait Images - CVPR 2020

Danger and Ethical Concerns

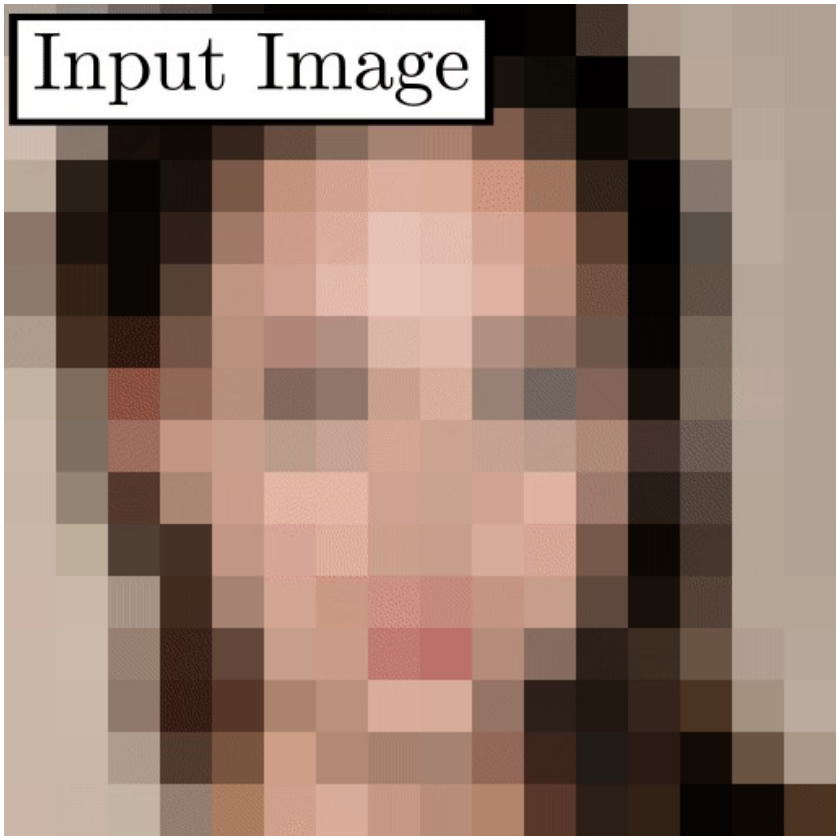


Image Super-resolution [PULSE, CVPR 2020]

Danger and Ethical Concerns

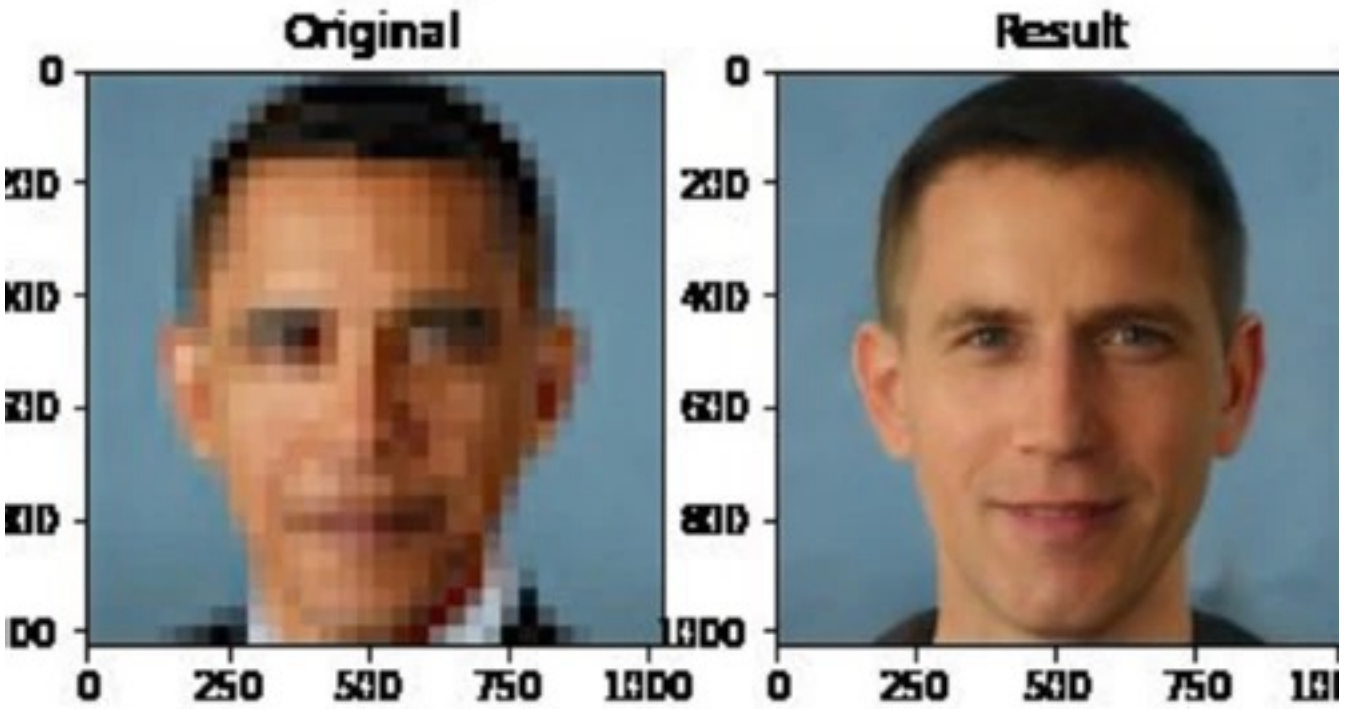
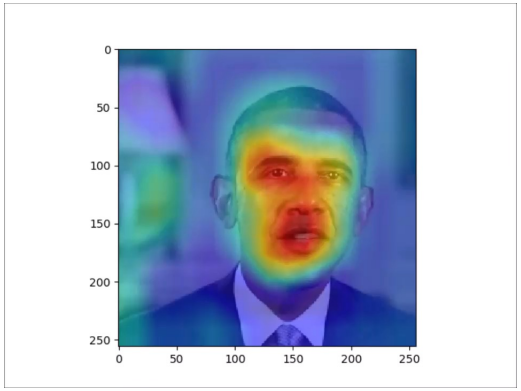


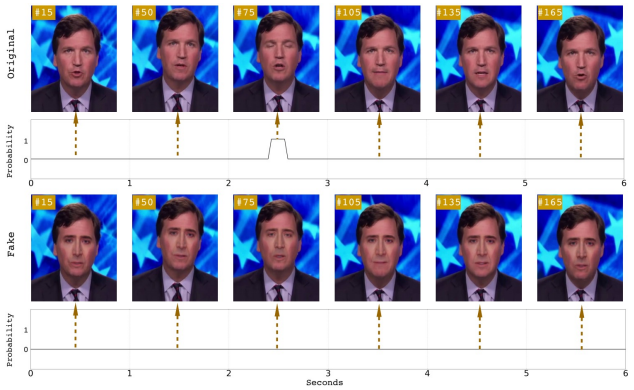
Image Super-resolution [PULSE, CVPR 2020]

Detecting Fake Faces



Ours

FaceForensics
Rössler et al., arXiv 2018



In Ictu Oculi
Li et al., arXiv 2018

Thank You!

16-726, Spring 2022

https://learning-image-synthesis.github.io/sp22/